

# A Logic of Knowing Why

Chao Xu and Yanjing Wang  
`{c.xu, y.wang}@pku.edu.cn`

Department of Philosophy, Peking University

**Abstract** When we say “I know why he was late”, we know not only the fact that he was late, but also an explanation of this fact. We propose a logical framework of “knowing why” inspired by the existing formal studies on why-questions, scientific explanation, and justification logic. We introduce the  $\mathcal{K}y_i$  operator into the language of epistemic logic to express “agent  $i$  knows why  $\varphi$ ” and propose a Kripke-style semantics of such expressions in terms of knowing an explanation of  $\varphi$ . We obtain two sound and complete axiomatizations w.r.t. two different model classes.

**Key words:** knowing why, why-questions, scientific explanation, epistemic logic, non-normal modal logic, axiomatization

## 1 Introduction

Ever since the seminal work by Hintikka [15], epistemic logic has grown into a major subfield of philosophical logic, which has unexpected applications in other fields such as computer science, AI, and game theory (cf. the handbook [29]). Standard epistemic logic focuses on propositional knowledge expressed by “knowing that  $\varphi$ ”. However, there are various knowledge expressions in terms of “knowing whether”, “knowing what”, “knowing how”, and so on, which have attracted a growing interest in recent years (cf. the survey [32]).

Among those “knowing-wh”,<sup>1</sup> “knowing why” is perhaps the most important driving force behind our advances in understanding the world and each other. For example, we may want to know why ([5]):

- the window is broken.
- the lump of potassium dissolved.
- he stayed in the café all day.
- cheetahs can run at high speeds.
- blood circulates in the body.

---

<sup>1</sup> Wh stands for the wh question words.

Intuitively, each “knowing why” expression corresponds to an embedded why-question. To some extent, the process of knowing the world is to answer why-questions about the world [16]. In fact, there is a very general connection between knowledge and wh-questions discovered by Hintikka in the framework of quantified epistemic logic [17]. For example, consider the question  $Q$  : “Who murdered Mary?”:

- The *presupposition* of  $Q$  is that the questioner knows that Mary was murdered by someone, formalized by  $\mathcal{K}\exists xM(x, \text{Mary})$ .
- The *desideratum* of  $Q$  is that the questioner knows who murdered Mary, which is formalized by  $\exists x\mathcal{K}M(x, \text{Mary})$ . The distinction between the desideratum and the presupposition highlights the difference between *de re* and *de dicto* readings of knowing who.
- One possible answer to  $Q$  is “John murdered Mary” formalized as  $M(\text{John}, \text{Mary})$ . However, telling the questioner this fact may not be enough to let the questioner know who murdered Mary since he or she may not have any idea on who John is. Therefore Hintikka also requires the following extra condition.
- *Conclusiveness* of the above answer also requires the questioner knows who John is ( $\exists x\mathcal{K}(x = \text{John})$ ). Conclusive answers realize the desideratum.

However, Hintikka viewed why-questions, such as  $Q$ : “Why  $\varphi$  is the case?”, as a special degenerated case where the presupposition and desideratum are the same:

- The *presupposition* of  $Q$  is  $\mathcal{K}\varphi$ ;
- The *desideratum* of  $Q$  is  $\mathcal{K}\varphi$ .

Hintikka then developed a different logical theory of why-questions in [18] using inquiry model and interpolation theorem of first-order logic. However, we do not think why-questions are special if we can quantify over the possible answers to them. Intuitively, an answer to a question “Why  $\varphi$ ?” is an explanation of the fact  $\varphi$ . In this paper, we take the view shared by Koura [20] and Schurz [26]:

- The *presupposition* of  $Q$  is that the questioner knows that there is an explanation for the fact  $\varphi$ :  $\mathcal{K}\exists xE(x, \varphi)$ .
- The *desideratum* of  $Q$  is the questioner knows why  $\varphi$ :  $\exists x\mathcal{K}E(x, \varphi)$ .

Note that if explanations are *factive*  $\exists xE(x, \varphi) \rightarrow \varphi$ , then the presupposition  $\mathcal{K}\exists xE(x, \varphi)$  also implies  $\mathcal{K}\varphi$  in quantified (normal) modal logic.

Now we have a preliminary logical form of knowing why in terms of the desideratum  $\exists x \mathcal{K}E(x, \varphi)$  of the corresponding why-question. The next questions are:

1. What are (good) *explanations*?
2. How can we capture the relation ( $E$  above) between an explanation and a proposition in logic?

The two questions are clearly related. To answer the first one, let us look back at the examples we mentioned at the beginning of this introduction. In fact there are different kinds of explanations [5]:

- Causal: The window broke because the stone was thrown at it.
- Nomic:<sup>2</sup> The lump of potassium dissolved since as a law of nature potassium reacts with water to form a soluble hydroxide.
- Psychological: He stayed in the café all day hoping to see her again.
- Darwinian: Cheetahs can run at high speeds because of the selective advantage this gives them in catching their prey.
- Functional: Blood circulates in order to supply the various parts of the body with oxygen and nutrients.

In philosophy of science, the emphasis is on *scientific explanations* to why questions, which mainly involve Nomic and Causal explanations in the above categorization [6,20,30]. According to Schurz [25] there are three major paradigms in understanding (scientific) explanations:<sup>3</sup>

- The *nomic expectability approach* initiated by Hempel [13], where a good explanation to  $\varphi$  should make the explanandum  $\varphi$  predictable or increases  $\varphi$ 's expectability.
- The *causality approach* (cf. e.g., [23]), where an explanation to  $\varphi$  should give a complete list of causes or relevant factors to  $\varphi$ .
- The *unification approach* (cf. e.g., [19]) where the focus is on the global feature of explanations in a coherent picture.

Our initial inspiration comes from the deductive-nomological model proposed by Hempel [14] in the first approach mentioned above, which is the mostly discussed (and criticized) model of explanation. The basic idea is that an explanation is a *derivation* of the explanandum from some universally quantified laws and some singular sentences. Although such a logical empiricistic approach arouse debates for decades,<sup>4</sup> it draws our

---

<sup>2</sup> Nomic explanations are explanation in terms of laws of nature.

<sup>3</sup> There are also various dimensions of each paradigm, e.g., probabilistic vs. non-probabilistic, singular events or laws to be explained.

<sup>4</sup> See [24,34] for critical surveys.

attention to the inner structure of explanations and its similarity to derivations in logic. In this paper, as the first step toward a logic of knowing why, we would like to stay neutral on different types of explanations and their models, and focus on the most abstract logical structure of (scientific) explanations. From a structuralist point of view, we only need to know how explanations compose and interact with each other without saying what they are exactly.

Now, as for the second question, how can we capture the explanatory relation between explanations and propositions in logic? Our next crucial inspiration came from *Justification Logic* proposed by Artemov [3]. Aiming at making up the gap between epistemic logic and the mainstream epistemology where justified true belief is the necessary basis of knowledge, justification logics are introduced based on the ideas of *Logic of Proof* (LP) [1].<sup>5</sup> Justification logic introduces formulas in the shape of  $t:\varphi$  into the logical language, read as “ $t$  is a justification of  $\varphi$ ”. Therefore, in justification logic we can talk about knowledge with an *explicit* justification. Moreover, justifications can be composed using various operations. For example, it is an axiom in the standard justification logic that  $t:(\varphi \rightarrow \psi) \wedge s:\varphi \rightarrow (t \cdot s):\psi$  where  $\cdot$  is the application operation of two justifications. Note that if we read  $t:\varphi$  as “ $t$  is an explanation of the fact  $\varphi$ ”, then this axiom also makes sense in general.

On the other hand, conceptually, justifications are quite different from explanations. For example, the fact that the shadow of a flagpole is  $x$  meters long may justify that the length of the pole is  $y$  meters given the specific time and location on earth. However, the length of the shadow of a flagpole clearly does not explain why the pole is  $y$  meters long, if we are looking for causal explanations. In general, a justification of  $\varphi$  gives a reason to *believing*  $\varphi$  (though not necessarily true), but an explanation gives a reason to *being*  $\varphi$ , presupposing the truth of  $\varphi$ . In this paper, we only make use of some technical apparatus of justification logic, and there are quite some differences in our framework compared to justification logic, which will be discussed in Section 4.

Now, putting all the above ideas together, we are almost ready to lay out the basis of our logic of knowing why. Following [32], we enrich the standard (multi-agent) epistemic language with a new “knowing why” operator  $\mathcal{K}y_i$ , instead of using a quantified modal language. Roughly speaking,  $\mathcal{K}y_i\varphi$  is essentially  $\exists t\mathcal{K}_i(t:\varphi)$ , although we do not allow quan-

---

<sup>5</sup> LP was invented to give an arithmetic semantics to intuitionistic logic under the Brouwer-Heyting-Kolmogorov provability interpretation.

tifiers and terms in the logical language.<sup>6</sup> As in [33,31], this will help us to control the expressive power of the logic in hopes of a computationally well-behaved logic. The semantics is based on the idea of Fitting model for justification logic. We will have a detailed comparison with justification in Section 4.

Since the language has both the standard epistemic operator  $\mathcal{K}_i$  and also the new “knowing why” modality  $\mathcal{K}y_i$ , there are lots of interesting things that can be expressed. For example,

- $\mathcal{K}_i p \wedge \neg \mathcal{K}y_i p$ , e.g., I know that Fermat’s last theorem is true but I do not know why.
- $\neg \mathcal{K}y_i p \wedge \mathcal{K}_i \mathcal{K}y_j p$ , e.g., I do not know why Fermat’s last theorem holds but I know that Andrew Wiles knows why.
- $\mathcal{K}_i \mathcal{K}_j p \wedge \neg \mathcal{K}y_i \mathcal{K}_j p$ , e.g., I know that you know that the paper has been accepted, but I do not know why you know.
- $\mathcal{K}y_i \mathcal{K}_j p \wedge \mathcal{K}_i \neg \mathcal{K}y_j p$ , e.g., I know why you know that the paper has been rejected, but I am sure you do not know why.

As we will see later, these situations are all satisfiable in our models.<sup>7</sup>

Before going into the technical details, it is helpful to summarize the aforementioned ideas:

- The language is inspired by the treatments of the logics of “knowing what”, and “knowing how”, where new modalities of such constructions are introduced, without using the full language of quantified epistemic logic.
- The formal treatment of explanations is inspired by the formal account of justifications in justification logics.
- The semantics of  $\mathcal{K}y_i$  is inspired by Hintikka’s logical formulation of the desideratum of Wh-questions:  $\exists t \mathcal{K}_i(t: \varphi)$ .

In the rest of the paper, Section 2 lays out the language, semantics and two proof systems of our knowing why logic; Section 3 proves the completeness of the two systems; Section 4 gives a detailed comparison with various versions of justification logic; Section 5 concludes the paper with philosophical discussions and future directions.

---

<sup>6</sup> Fitting proposed an quantified justification logic in [9], and discussed briefly what can be expressed if the language also includes the normal  $\mathcal{K}$  operator. Our language can then be viewed as a fragment of this quantified justification logic extended with  $\mathcal{K}$ .

<sup>7</sup> According to our semantics to be introduced later, it is also allowed to know why for different reasons (for different people), which can help to model mutual misunderstanding.

## 2 A Logic of Knowing Why

**Definition 1 (Language ELKy)** Given a countable set  $\mathbf{I}$  of agent names and a countably infinite set  $\mathbf{P}$  of basic propositional letters, the language of ELKy is defined as

$$\varphi ::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \psi) \mid \mathcal{K}_i\varphi \mid \mathcal{K}y_i\varphi$$

where  $p \in \mathbf{P}$  and  $i \in \mathbf{I}$ .

We use standard abbreviations for  $\perp$ ,  $\varphi \rightarrow \psi$ ,  $\varphi \vee \psi$ , and  $\widehat{\mathcal{K}}_i\varphi$  (the dual of  $\mathcal{K}_i\varphi$ ).  $\mathcal{K}y_i\phi$  says that agent  $i$  knows why  $\phi$  (is the case).

Intuitively, necessitation rule for  $\mathcal{K}y_i$  should not hold, e.g., although something is a tautology, you may not know why it is a tautology. Borrowing the idea from justification logic, we introduce a special set of “self-evident” tautologies which the agents are assumed to know why. Please see Section 4 for the comparison with *constant specifications* in justification logic where all the axioms in the logic are included.

**Definition 2 (Tautology Ground  $\Lambda$ )** Tautology Ground  $\Lambda$  is a set of propositional tautologies.

For example,  $\Lambda$  can be the set of all the instances of  $\varphi \wedge \psi \rightarrow \varphi$  and  $\varphi \wedge \psi \rightarrow \psi$ . As we will see later, under such a  $\Lambda$ ,  $\mathcal{K}y_i(\varphi \wedge \psi \rightarrow \varphi)$  will be valid, which helps the agents to reason more.

The model of our language ELKy is similar to the Fitting model of justification logic [8]. Note that we do not have the justification terms in the logical language, but we do have a set  $E$  of explanations as semantic objects in the models. In this work, we require the accessibility relation to be equivalence relations to accommodate the S5 epistemic logic.

**Definition 3 (ELKy-Model)** An ELKy-model  $\mathcal{M}$  is a tuple  $(W, E, \{R_i \mid i \in \mathbf{I}\}, \mathcal{E}, V)$  where:

- $W$  is a non-empty set of possible worlds.
- $E$  is a non-empty set of explanations satisfying the following conditions
  - (a) If  $s, t \in E$ , then a new explanation  $(s \cdot t) \in E$ ;
  - (b) A special symbol  $e$  is in  $E$ .
- $R_i \subseteq W \times W$  is an equivalence relation over  $W$ .
- $\mathcal{E} : E \times \mathbf{ELKy} \rightarrow 2^W$  is an admissible explanation function satisfying the following conditions:
  - (I)  $\mathcal{E}(s, \varphi \rightarrow \psi) \cap \mathcal{E}(t, \varphi) \subseteq \mathcal{E}(s \cdot t, \psi)$ .
  - (II) If  $\varphi \in \Lambda$ , then  $\mathcal{E}(e, \varphi) = W$ .

- $V : \mathbf{P} \rightarrow 2^W$  is a valuation function.

Note that  $E$  does not depend on possible worlds, thus it can be viewed as a *constant domain* of explanations closed under an *application* operator  $\cdot$  which combines two explanations into one. The special element  $e$  in  $E$  is the *self-evident explanation* which is uniform for all the self-evident formulas in  $\Lambda$ . The admissible explanation function  $\mathcal{E}$  specifies the set of worlds where  $t$  is an explanation of  $\varphi$ . It is possible that some formula has *no* explanation on some world, and some formula has more than one explanation on some world, e.g., one theorem may have different proofs. The first condition of  $\mathcal{E}$  captures the composition of explanations resembling the reasoning of knowing why by *modus ponens*, which amounts to the later axiom  $\mathcal{K}y_i(\varphi \rightarrow \psi) \rightarrow (\mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\psi)$ .

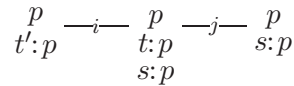
#### Definition 4 (Semantics)

The satisfaction relation of **ELKy** formulas on pointed models is as below:

$\mathcal{M}, w \models \top$	<i>always</i>
$\mathcal{M}, w \models p$	$\iff w \in V(p)$
$\mathcal{M}, w \models \neg\varphi$	$\iff \mathcal{M}, w \not\models \varphi$
$\mathcal{M}, w \models \varphi \wedge \psi$	$\iff \mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
$\mathcal{M}, w \models \mathcal{K}_i\varphi$	$\iff \mathcal{M}, v \models \varphi$ for each $v$ such that $wR_iv$ .
$\mathcal{M}, w \models \mathcal{K}y_i\varphi$	$\iff$ (1) $\exists t \in E$ , for each $v$ such that $wR_iv$ , $v \in \mathcal{E}(t, \varphi)$ ; (2) $\forall v \in W$ , $wR_iv$ , $v \models \varphi$ .

Now it is clear that our  $\mathcal{K}y_i\varphi$  is roughly  $\exists t\mathcal{K}_i(t:\varphi) \wedge \mathcal{K}_i\varphi$  though there are subtle details to be discussed in Section 4 when compared to justification logic. Also note that  $\mathcal{K}y_i\varphi \rightarrow \mathcal{K}_i\varphi$  is clearly valid, but the  $\mathcal{K}y_i$ -necessitation is not, since not all the valid formulas are explained except those in  $\Lambda$ . Moreover, things we usually take for granted are not valid either, e.g.,  $\mathcal{K}y_i\varphi \wedge \mathcal{K}y_i\psi \rightarrow \mathcal{K}y_i(\varphi \wedge \psi)$  is not valid in general: I have explanations  $\varphi$  and  $\psi$  respectively does not mean I have an explanation for the co-occurrence of the two, e.g., quantum mechanics and general relativity have their own explanatory power on microcosm and macrocosm respectively, but a “theory of everything” is not obtained by simply putting these two theories together.

As an example, in the following model (reflexive arrows are omitted),  $\mathcal{K}_ip \wedge \neg\mathcal{K}y_ip \wedge \mathcal{K}y_jp \wedge \mathcal{K}_i\mathcal{K}y_jp$  holds on the middle world.



In this paper, we also consider models with special properties. First of all, we are interested in the models where explanations are always correct,<sup>8</sup> i.e., if a proposition has an explanation on a world, then it must be true.

**Definition 5 (Factivity Property)** *An ELKy-model  $\mathcal{M}$  has factivity property provided that, whenever  $w \in \mathcal{E}(t, \varphi)$ , then  $\mathcal{M}, w \models \varphi$ .*

Besides factivity, it is also debatable whether knowing why is introspective, i.e., are the following reasonable? Note that they are not valid without further conditions on the models.

$$\begin{aligned} \mathcal{K}_i\varphi &\rightarrow \mathcal{K}y_i\mathcal{K}_i\varphi, \neg\mathcal{K}_i\varphi \rightarrow \mathcal{K}y_i\neg\mathcal{K}_i\varphi, \\ \mathcal{K}y_i\varphi &\rightarrow \mathcal{K}y_i\mathcal{K}y_i\varphi, \neg\mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\neg\mathcal{K}y_i\varphi \end{aligned}$$

One may argue that there is always a self-evident explanation to your own knowledge or ignorance, but another may say it happens a lot that you just forgot why you know some facts. Things can be even more complicated regarding nested  $\mathcal{K}y_i$ . Your explanation for why  $\phi$  holds may be quite different from the explanation for why you know why  $\phi$ , e.g., the window is broken ( $\phi$ ) because you know a stone was thrown at it, and you know why  $\phi$  because someone told you so. On the other hand, if you know why a theorem holds because of a proof, it seems reasonable to assume that you know why you know why the theorem holds: you can just verify the proof. The cases of negative introspection may invoke more debates.

As a first attempt to a logic of knowing why, we want to remain neutral in the philosophical debate, but would like to make it technically possible to handle the cases when introspection is considered reasonable. The following property guarantees that the above introspection axioms are valid.

**Definition 6 (Introspection Property)** *An ELKy-model  $\mathcal{M}$  has introspection property provided that, whenever  $\mathcal{M}, w \models \varphi$  and  $\varphi$  has the form of  $\mathcal{K}_i\psi$  or  $\neg\mathcal{K}_i\psi$  or  $\mathcal{K}y_i\psi$  or  $\neg\mathcal{K}y_i\psi$ , then  $\exists t \in E$ , for each  $v$  such that  $wR_tv$ ,  $v \in \mathcal{E}(t, \varphi)$ .*

We use  $\mathbb{C}$ ,  $\mathbb{C}_F$ ,  $\mathbb{C}_I$ ,  $\mathbb{C}_{FI}$  to denote respectively the model classes of all ELKy-models, factive models, introspective models, and models with both properties. Obviously, we have  $\mathbb{C}_F \subseteq \mathbb{C}$ ,  $\mathbb{C}_I \subseteq \mathbb{C}$ ,  $\mathbb{C}_{FI} \subseteq \mathbb{C}_F$ , and  $\mathbb{C}_{FI} \subseteq \mathbb{C}_I$ . In the following, we write  $\Gamma \models_{\mathbb{C}} \varphi$  if  $\mathcal{M}, w \models \Gamma$  implies  $\mathcal{M}, w \models \varphi$ , for any  $\mathcal{M} \in \mathbb{C}$  and any  $w$  in  $\mathcal{M}$ . Similar for  $\mathbb{C}_F$ ,  $\mathbb{C}_I$ ,  $\mathbb{C}_{FI}$ .

<sup>8</sup> The corresponding factivity  $t:\varphi \rightarrow \varphi$  in justification logic is guaranteed by the reflexivity of the models. See the later discussion in Section 4 on the Fitting semantics.



However, as we will see below, factivity does not affect the valid formulas. For an arbitrary  $\mathcal{M} \in \mathbb{C}$ , we can construct a new **ELKy**-model  $\mathcal{M}^F \in \mathbb{C}_F$  which has factivity. Given  $\mathcal{M} = (W, E, \{R_i \mid i \in \mathbf{I}\}, \mathcal{E}, V)$ , let  $\mathcal{M}^F = (W, E, \{R_i \mid i \in \mathbf{I}\}, \mathcal{E}^F, V)$  where:

$$\mathcal{E}^F(t, \varphi) = \mathcal{E}(t, \varphi) - \{u \mid \mathcal{M}, u \not\models \varphi\}$$

We will show that  $\mathcal{M}, w$  and  $\mathcal{M}^F, w$  satisfy the same **ELKy** formulas, thus by the above definition of  $\mathcal{E}^F$ , it is clear that  $\mathcal{M}^F$  has factivity.

**Lemma 7** *For any **ELKy**-formula  $\varphi$  and any  $w \in W$ ,  $\mathcal{M}, w \models \varphi$  if and only if  $\mathcal{M}^F, w \models \varphi$ .*

**PROOF** We can prove it by induction on the structure of formulas. It is trivial for the atomic, boolean, and  $\mathcal{K}\psi$  cases since  $\mathcal{M}^F$  only differs from  $\mathcal{M}$  in  $\mathcal{E}^F$ . We just need to prove that  $\mathcal{M}, w \models \mathcal{K}y_i\psi$  iff  $\mathcal{M}^F, w \models \mathcal{K}y_i\psi$ .

- $\Rightarrow$  Suppose  $\mathcal{M}, w \models \mathcal{K}y_i\psi$ . Then  $\exists t \in E$ , for each  $v$  such that  $wR_iv$ ,  $v \in \mathcal{E}(t, \psi)$  and  $v \models \psi$ . Thus  $v \notin \{u \mid \mathcal{M}, u \not\models \psi\}$ . Therefore we have  $v \in \mathcal{E}^F(t, \psi)$ . Hence by IH we have  $\mathcal{M}^F, w \models \mathcal{K}y_i\psi$ .
- $\Leftarrow$  Suppose  $\mathcal{M}^F, w \models \mathcal{K}y_i\psi$ . Then  $\exists t \in E$ , for each  $v$  such that  $wR_iv$ ,  $v \in \mathcal{E}^F(t, \psi)$  and  $v \models \psi$ . By the definition of  $\mathcal{E}^F$ , we still have  $v \in \mathcal{E}(t, \psi)$ . Hence by IH  $\mathcal{M}, w \models \mathcal{K}y_i\psi$ .

□

**Theorem 8** *For any set  $\Gamma \cup \{\varphi\}$  of formulas,  $\Gamma \models_{\mathbb{C}} \varphi$  if and only if  $\Gamma \models_{\mathbb{C}_F} \varphi$ .*

**PROOF**

- $\Rightarrow$  Suppose  $\Gamma \models_{\mathbb{C}} \varphi$  and  $\Gamma \not\models_{\mathbb{C}_F} \varphi$ . By  $\Gamma \not\models_{\mathbb{C}_F} \varphi$ , there exists a factive model  $\mathcal{N} \in \mathbb{C}_F$  such that  $\mathcal{N}, w \models \Gamma$  and  $\mathcal{N}, w \not\models \varphi$  for some  $w$  in  $\mathcal{N}$ . Since  $\mathbb{C}_F \subseteq \mathbb{C}$ , we have  $\mathcal{N} \in \mathbb{C}$ . Thus  $\Gamma \not\models_{\mathbb{C}} \varphi$ . Contradiction.
- $\Leftarrow$  Suppose  $\Gamma \models_{\mathbb{C}_F} \varphi$  and  $\Gamma \not\models_{\mathbb{C}} \varphi$ . Then there exists a model  $\mathcal{M} \in \mathbb{C}$  such that  $\mathcal{M} \models \Gamma$  and  $\mathcal{M} \not\models \varphi$ . By lemma 7, we can construct an  $\mathcal{M}^F \in \mathbb{C}_F$  such that  $\mathcal{M}^F \models \Gamma$  and  $\mathcal{M}^F \not\models \varphi$ . Thus  $\Gamma \not\models_{\mathbb{C}_F} \varphi$ . Contradiction.

□

Now we consider the introspective models.

**Lemma 9** *If  $\mathcal{M}$  is introspective, then so is  $\mathcal{M}^F$ .*

**PROOF** Suppose  $\mathcal{M}^F, w \models \varphi$  and  $\varphi$  has the form of  $\mathcal{K}_i\psi$  or  $\neg\mathcal{K}_i\psi$  or  $\mathcal{K}y_i\psi$  or  $\neg\mathcal{K}y_i\psi$ . By lemma 7, we have  $\mathcal{M}, w \models \varphi$ . Since  $\mathcal{M}$  has introspection

property, we have that  $\exists t \in E$ , for each  $v$  such that  $wR_iv$ ,  $v \in \mathcal{E}(t, \varphi)$ . Since  $\mathcal{M}^F, w \models \varphi$  and  $R_i$  is an equivalence relation, we have  $\mathcal{M}^F, v \models \varphi$  for each  $v$  such that  $wR_iv$ . Thus  $v \notin \{u \mid \mathcal{M}, u \not\models \varphi\}$ . Thus  $v \in \mathcal{E}^F(t, \varphi)$ . Hence  $\exists t \in E$ , for each  $v$  such that  $wR_iv$ ,  $v \in \mathcal{E}^F(t, \varphi)$ . Therefore  $\mathcal{M}^F$  has introspection property.  $\square$

It is then easy to show:

**Theorem 10** For any set  $\Gamma \cup \{\varphi\}$ ,  $\Gamma \models_{\mathbb{C}_I} \varphi$  if and only if  $\Gamma \models_{\mathbb{C}_{FI}} \varphi$ .

Theorems 8 and 10 showed that factivity is neglectable w.r.t. the logic.

In the following, we present two proof systems which differ only on the introspection axioms of  $\mathcal{K}_{y_i}$  essentially. In the next section, we will show their completeness w.r.t.  $\mathbb{C}$  and  $\mathbb{C}_I$  respectively.

#### System SKY

TAUT Classical Propositional Axioms

DISTK  $\mathcal{K}_i(\varphi \rightarrow \psi) \rightarrow (\mathcal{K}_i\varphi \rightarrow \mathcal{K}_i\psi)$

DISTY  $\mathcal{K}_{y_i}(\varphi \rightarrow \psi) \rightarrow (\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_{y_i}\psi)$

MP Modus Ponens

T  $\mathcal{K}_i\varphi \rightarrow \varphi$

NECK  $\vdash \varphi \Rightarrow \vdash \mathcal{K}_i\varphi$

4  $\mathcal{K}_i\varphi \rightarrow \mathcal{K}_i\mathcal{K}_i\varphi$

NECKY If  $\varphi \in A$ , then  $\vdash \mathcal{K}_{y_i}\varphi$

5  $\neg\mathcal{K}_i\varphi \rightarrow \mathcal{K}_i\neg\mathcal{K}_i\varphi$

PRES  $\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\varphi$

4YK  $\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\mathcal{K}_{y_i}\varphi$

PRES is the presupposition axiom which says “knowing that” is necessary for “knowing why”. 4YK is the positive introspection of “knowing why” by “knowing that”.<sup>9</sup> The reader may wonder about the corresponding negative introspection of 4YK and it is provable in SKY.

**Proposition 1.** 5YK:  $\neg\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\neg\mathcal{K}_{y_i}\varphi$  is provable in SKY.

PROOF

(1)  $\mathcal{K}_i\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_{y_i}\varphi$

T

(2)  $\neg\mathcal{K}_{y_i}\varphi \rightarrow \neg\mathcal{K}_i\mathcal{K}_{y_i}\varphi$

Contraposition (1)

(3)  $\neg\mathcal{K}_i\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\neg\mathcal{K}_i\mathcal{K}_{y_i}\varphi$

5

(4)  $\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\mathcal{K}_{y_i}\varphi$

4

(5)  $\neg\mathcal{K}_i\mathcal{K}_{y_i}\varphi \rightarrow \neg\mathcal{K}_{y_i}\varphi$

Contraposition (4)

$\square$

(6)  $\mathcal{K}_i(\neg\mathcal{K}_i\mathcal{K}_{y_i}\varphi \rightarrow \neg\mathcal{K}_{y_i}\varphi)$

NECK(5)

(7)  $\mathcal{K}_i\neg\mathcal{K}_i\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\neg\mathcal{K}_{y_i}\varphi$

MP(6), DISTK

(8)  $\neg\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\neg\mathcal{K}_i\mathcal{K}_{y_i}\varphi$

MP(2)(3)

(9)  $\neg\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\neg\mathcal{K}_{y_i}\varphi$

MP(7)(8)

<sup>9</sup> Note that this is not one of the four introspection axioms of  $\mathcal{K}_{y_i}$  mentioned earlier.

Note that the choice of  $\Lambda$  and NECKY in SKY also give us some flexibility in the logic.

System SKYI is obtained by replacing 4, 5 and 4YK in SKY by the those four stronger introspection axioms of  $\mathcal{K}y_i$ :

$$\begin{array}{ll} 4KY \quad \mathcal{K}_i\varphi \rightarrow \mathcal{K}y_i\mathcal{K}_i\varphi & 4Y \quad \mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\mathcal{K}y_i\varphi \\ 5KY \quad \neg\mathcal{K}_i\varphi \rightarrow \mathcal{K}y_i\neg\mathcal{K}_i\varphi & 5Y \quad \neg\mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\neg\mathcal{K}y_i\varphi \end{array}$$

It is straightforward to show that SKYI is deductively stronger than SKY.

**Proposition 11** *The following are provable in SKYI*

$$\begin{array}{ll} 4 \quad \mathcal{K}_i\varphi \rightarrow \mathcal{K}_i\mathcal{K}_i\varphi & 4YK \quad \mathcal{K}y_i\varphi \rightarrow \mathcal{K}_i\mathcal{K}y_i\varphi \\ 5 \quad \neg\mathcal{K}_i\varphi \rightarrow \mathcal{K}_i\neg\mathcal{K}_i\varphi & 5YK \quad \neg\mathcal{K}y_i\varphi \rightarrow \mathcal{K}_i\neg\mathcal{K}y_i\varphi \end{array}$$

### 3 Soundness and Completeness

Due to Theorems 8 and 10, we only need to prove soundness and completeness w.r.t.  $\mathbb{C}$  and  $\mathbb{C}_I$  instead of  $\mathbb{C}_F$  and  $\mathbb{C}_{FI}$  respectively.

**Theorem 12 (Soundness)** *SKY and SKYI are sound for  $\mathbb{C}$  and  $\mathbb{C}_I$  respectively.*

**PROOF** Since ELKy-models are based on S5 Kripke models, the standard axioms of system S5 are all valid. So we just need to check the rest. First we check the non-trivial axioms and rules of SKY on  $\mathbb{C}$ .

DISTY:  $\mathcal{K}y_i(\varphi \rightarrow \psi) \rightarrow (\mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\psi)$

Suppose  $w \models \mathcal{K}y_i(\varphi \rightarrow \psi)$  and  $w \models \mathcal{K}y_i\varphi$ . Then by the definition of  $\models$ ,  $\exists s, t \in E$ , for any  $v$  such that  $wR_iv, v \in \mathcal{E}(s, \varphi \rightarrow \psi), v \in \mathcal{E}(t, \varphi), v \models \varphi \rightarrow \psi$ , and  $v \models \varphi$ . Then we have  $v \models \psi$  and  $v \in \mathcal{E}(s, \varphi \rightarrow \psi) \cap \mathcal{E}(t, \varphi)$ . By the condition (I) of  $\mathcal{E}$ , we have  $v \in \mathcal{E}(s \cdot t, \psi)$ . Hence  $w \models \mathcal{K}y_i\psi$ .

PRES:  $\mathcal{K}y_i\varphi \rightarrow \mathcal{K}_i\varphi$

Suppose  $w \models \mathcal{K}y_i\varphi$ . Then for any  $v$  such that  $wR_iv$ , we have  $v \models \varphi$ . Thus  $w \models \mathcal{K}_i\varphi$ .

4YK:  $\mathcal{K}y_i\varphi \rightarrow \mathcal{K}_i\mathcal{K}y_i\varphi$

By the fact that the relations are equivalence relations.

NECKY Suppose  $\varphi \in \Lambda$ . Since  $\Lambda$  is a set of tautologies, thus we have  $\forall w \in W, w \models \varphi$ . By the condition (II) of  $\mathcal{E}$ ,  $\forall w \in W, \exists e \in E$ , for any  $v$  such that  $wR_iv, v \in \mathcal{E}(e, \varphi)$ . Therefore it follows that  $\models \mathcal{K}y_i\varphi$ . Hence NECKY is valid.

Validity of the introspection axioms of  $\mathbf{SKYI}$  on  $\mathbb{C}_I$  are trivial based on the introspective property and the fact that  $R_i$  is an equivalence relation.  $\square$

To establish completeness, we build a canonical model for each consistent set of  $\mathbf{ELKy}$  formulas. We will first show the completeness of  $\mathbf{SKY}$  over  $\mathbb{C}$ , and the completeness of  $\mathbf{SKYI}$  over  $\mathbb{C}_I$  is then straightforward.

Let  $\Omega$  be the set of all maximal  $\mathbf{SKY}$ -consistent sets of formulas. For any maximal consistent set (abbr. MCS)  $\Gamma$ , let  $\Gamma_i^\# = \{\mathcal{K}y_i\varphi \mid \mathcal{K}y_i\varphi \in \Gamma\} \cup \{\varphi \mid \mathcal{K}_i\varphi \in \Gamma\}$ .

**Definition 13 (Canonical model for  $\mathbf{SKY}$ )** *The canonical model  $\mathcal{M}^c$  for  $\mathbf{SKY}$  is a tuple  $(W^c, E^c, \{R_i^c \mid i \in \mathbf{I}\}, \mathcal{E}^c, V^c)$  where:*

- $E^c$  is defined in BNF:  $t ::= e \mid \varphi \mid (t \cdot t)$  where  $\varphi \in \mathbf{ELKy}$ .
- $W^c = \{\langle \Gamma, F, \{f_i \mid i \in \mathbf{I}\} \rangle \mid \langle \Gamma, F \rangle \in \Omega \times \mathcal{P}(E^c \times \mathbf{ELKy}), f_i : \{\varphi \mid \mathcal{K}y_i\varphi \in \Gamma\} \rightarrow E^c \text{ such that } F \text{ and } \vec{f} \text{ satisfy the conditions below}\}$ :
  - (i) If  $\langle s, \varphi \rightarrow \psi \rangle, \langle t, \varphi \rangle \in F$ , then  $\langle s \cdot t, \psi \rangle \in F$ ;
  - (ii) If  $\varphi \in \Lambda$ , then  $\langle e, \varphi \rangle \in F$ ;
  - (iii) For any  $i \in \mathbf{I}$ ,  $\mathcal{K}y_i\varphi \in \Gamma$  implies  $\langle f_i(\varphi), \varphi \rangle \in F$ .
- $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$  iff (1)  $\Gamma_i^\# \subseteq \Delta$ , and (2)  $f_i = g_i$ .
- $\mathcal{E}^c: E^c \times \mathbf{ELKy} \rightarrow 2^{W^c}$  defined by  $\mathcal{E}^c(t, \varphi) = \{\langle \Gamma, F, \vec{f} \rangle \mid \langle t, \varphi \rangle \in F\}$ .
- $V^c(p) = \{\langle \Gamma, F, \vec{f} \rangle \mid p \in \Gamma\}$ .

In the above we write  $\vec{f}$  for  $\{f_i \mid i \in \mathbf{I}\}$ . Essentially,  $f_i$  is a witness function picking one  $t$  for each formula in  $\{\varphi \mid \mathcal{K}y_i\varphi \in \Gamma\}$ . It can be used to construct the possible worlds for the existence lemma for  $\neg\mathcal{K}y_i\phi$ . We do need such witness functions for each  $i$ , since  $i, j$  can have different explanations for  $\varphi$ . In the definition of  $R_i^c$ , we need to make sure the selected witnesses are the same for  $i$ .

Now we need to show that the canonical model is well-defined:

- $\mathcal{E}^c$  satisfies conditions (I) and (II) in the definition of  $\mathbf{ELKy}$ -models.
- $R_i^c$  is an equivalence relation.
- $W^c$  is not empty. Actually, we will prove a stronger one: for any  $\Gamma \in \Omega$ , there exist  $F$  and  $\vec{f}$  such that  $\langle \Gamma, F, \vec{f} \rangle \in W^c$ .

**Proposition 14**  $\mathcal{E}^c$  satisfies the conditions (I) and (II) of  $\mathbf{ELKy}$ -models.

PROOF

- (1) Suppose  $\langle \Gamma, F, \vec{f} \rangle \in \mathcal{E}^c(s, \varphi \rightarrow \psi) \cap \mathcal{E}^c(t, \varphi)$ . By the definition of  $\mathcal{E}^c$ , we have  $\langle s, \varphi \rightarrow \psi \rangle, \langle t, \varphi \rangle \in F$ . By the condition (i) of  $F$  in the definition of  $W^c$ , we have  $\langle s \cdot t, \psi \rangle \in F$ . Hence it follows that  $\langle \Gamma, F, \vec{f} \rangle \in \mathcal{E}^c(s \cdot t, \psi)$ . Therefore  $\mathcal{E}^c(s, \varphi \rightarrow \psi) \cap \mathcal{E}^c(t, \varphi) \subseteq \mathcal{E}^c(s \cdot t, \psi)$ .

- (2) Suppose  $\varphi \in \Lambda$ . For an arbitrary  $\langle \Gamma, F, \vec{f} \rangle \in W^c$ , by condition (ii) in the definition of  $W^c$ , we have  $\langle e, \varphi \rangle \in F$ . By the definition of  $\mathcal{E}^c$ , we have  $\langle \Gamma, F, \vec{f} \rangle \in \mathcal{E}^c(e, \varphi)$ . Hence  $\mathcal{E}^c(e, \varphi) = W^c$ .

□

Before proceeding further, we prove the following handy proposition.

**Proposition 15** *If  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$ , then (1)  $\mathcal{K}_i \varphi \in \Gamma$  iff  $\mathcal{K}_i \varphi \in \Delta$  and (2)  $\mathcal{K}_i \varphi \in \Gamma$  iff  $\mathcal{K}_i \varphi \in \Delta$ .*

PROOF

- (1) Suppose  $\mathcal{K}_i \varphi \in \Gamma$ . By the definition of  $R_i^c$ , we have  $\mathcal{K}_i \varphi \in \Delta$ .  
 Suppose  $\mathcal{K}_i \varphi \in \Delta$  and  $\mathcal{K}_i \varphi \notin \Gamma$ . By the property of MCS, we have  $\neg \mathcal{K}_i \varphi \in \Gamma$ . By the provable 5YK ( $\neg \mathcal{K}_i \varphi \rightarrow \mathcal{K}_i \neg \mathcal{K}_i \varphi$ ) and the property of MCS, we have  $\mathcal{K}_i \neg \mathcal{K}_i \varphi \in \Gamma$ . By the definition of  $R_i^c$ , we have  $\neg \mathcal{K}_i \varphi \in \Delta$ . Contradiction.
- (2) Suppose  $\mathcal{K}_i \varphi \in \Gamma$ . By axiom 4 and the property of MCS, we have that  $\mathcal{K}_i \mathcal{K}_i \varphi \in \Gamma$ . By the definition of  $R_i^c$ , we have that  $\mathcal{K}_i \varphi \in \Delta$ .  
 Suppose  $\mathcal{K}_i \varphi \in \Delta$  and  $\mathcal{K}_i \varphi \notin \Gamma$ . By the property of MCS, we have that  $\neg \mathcal{K}_i \varphi \in \Gamma$ . By axiom 5 we have that  $\mathcal{K}_i \neg \mathcal{K}_i \varphi \in \Gamma$ . Then we have  $\neg \mathcal{K}_i \varphi \in \Delta$  by the definition of  $R_i^c$ . Contradiction.

□

**Proposition 16**  *$R_i^c$  is an equivalence relation.*

PROOF We just need to prove  $R_i^c$  is reflexive, transitive, and symmetric.

- (1)  $R_i^c$  is reflexive: For all  $\mathcal{K}_i \varphi, \mathcal{K}_i \psi \in \Gamma$ , by axiom T we have  $\psi \in \Gamma$ . Hence we have  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Gamma, F, \vec{f} \rangle$  by the definition of  $R_i^c$ .
- (2)  $R_i^c$  is transitive: Suppose  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$  and  $\langle \Delta, G, \vec{g} \rangle R_i^c \langle \Theta, H, \vec{h} \rangle$ . Suppose  $\mathcal{K}_i \varphi, \mathcal{K}_i \psi \in \Gamma$ . By the definition of  $R_i^c$ , we have  $f_i = g_i = h_i$ . By Proposition 15, we have  $\mathcal{K}_i \varphi, \mathcal{K}_i \psi \in \Delta \cap \Theta$ . Then by axiom T, we have  $\mathcal{K}_i \varphi, \psi \in \Theta$ . Therefore  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Theta, H, \vec{h} \rangle$  by the definition of  $R_i^c$ .
- (3)  $R_i^c$  is symmetric: Suppose  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$ . Then we have  $f_i = g_i$ . Suppose  $\mathcal{K}_i \varphi, \mathcal{K}_i \psi \in \Delta$ . By proposition 15, we have  $\mathcal{K}_i \varphi \in \Gamma$  and  $\mathcal{K}_i \psi \in \Gamma$ . By axiom T,  $\psi \in \Gamma$ , thus  $\langle \Delta, G, \vec{g} \rangle R_i^c \langle \Gamma, F, \vec{f} \rangle$ .

□

To prove that for any  $\Gamma \in \Omega$ , there exist  $F$  and  $\vec{f}$  such that  $\langle \Gamma, F, \vec{f} \rangle \in W^c$ , we define the following construction.

**Definition 17** Given any  $\Gamma \in \Omega$ , construct  $F^\Gamma$  and  $\vec{f}^\Gamma$  as follows:

- $F_0^\Gamma = \{\langle \varphi, \varphi \rangle \mid \exists i \in \mathbf{I}, \mathcal{K}y_i \varphi \in \Gamma\} \cup \{\langle e, \varphi \rangle \mid \varphi \in \Lambda\}$
- $F_{n+1}^\Gamma = F_n^\Gamma \cup \{\langle s \cdot t, \psi \rangle \mid \langle s, \varphi \rightarrow \psi \rangle \in F_n^\Gamma, \langle t, \varphi \rangle \in F_n^\Gamma\}$
- $F^\Gamma = \bigcup_{n \in \mathbb{N}} F_n^\Gamma$ .
- $\forall i \in \mathbf{I}, f_i^\Gamma : \{\varphi \mid \mathcal{K}y_i \varphi \in \Gamma\} \rightarrow E^c, f_i^\Gamma(\varphi) = \varphi$ .

By the construction of  $F_n^\Gamma (n \in \mathbb{N})$ ,  $\{F_n^\Gamma \mid n \in \mathbb{N}\}$  is monotonic. i.e.,  $\forall m, n \in \mathbb{N}$ , if  $m \leq n$ , then  $F_m^\Gamma \subseteq F_n^\Gamma$ .

**Proposition 18** For any  $\Gamma \in \Omega$ ,  $\langle \Gamma, F^\Gamma, \vec{f}^\Gamma \rangle \in W^c$ .

PROOF To prove  $\langle \Gamma, F^\Gamma, \vec{f}^\Gamma \rangle \in W^c$ , we just need to show that  $F^\Gamma$  satisfies conditions (i)-(iii) in the definition of  $W^c$ .

- Suppose  $\langle s, \varphi \rightarrow \psi \rangle, \langle t, \varphi \rangle \in F^\Gamma$ . By monotonicity of  $\{F_n^\Gamma \mid n \in \mathbb{N}\}$ , there exists  $k \in \mathbb{N}$  such that  $\langle s, \varphi \rightarrow \psi \rangle, \langle t, \varphi \rangle \in F_k^\Gamma$ . Thus we have  $\langle s \cdot t, \psi \rangle \in F_{k+1}^\Gamma$  by the construction of  $F_k^\Gamma (k \in \mathbb{N})$ . Hence  $\langle s \cdot t, \psi \rangle \in F^\Gamma$ , thus  $F^\Gamma$  satisfies condition (i).
- Suppose  $\varphi \in \Lambda$ . By the construction of  $F_0^\Gamma$ , we have  $\langle e, \varphi \rangle \in F_0^\Gamma$ . Thus  $\langle e, \varphi \rangle \in F^\Gamma$ . Hence  $F$  satisfies condition (ii).
- Suppose  $\mathcal{K}y_i \varphi \in \Gamma$ . Then we have  $\langle \varphi, \varphi \rangle \in F^\Gamma$  by the construction of  $F_0^\Gamma$  and  $F^\Gamma$ . Since  $\mathcal{K}y_i \varphi \in \Gamma$ , by the construction of  $f_i^\Gamma$ , we have  $\varphi \in \text{dom}(f_i^\Gamma)$  and  $f_i^\Gamma(\varphi) = \varphi$ . Thus we have  $\langle f_i^\Gamma(\varphi), \varphi \rangle \in F^\Gamma$ . Hence, we have that  $F^\Gamma$  and  $\vec{f}^\Gamma$  satisfy condition (iii).

□

This completes the proof to show  $\mathcal{M}^c$  is well-defined. Now we can establish the existence lemmas for  $\mathcal{K}_i$  and  $\mathcal{K}y_i$ .

**Lemma 19 ( $\mathcal{K}_i$  Existence Lemma)** For any  $\langle \Gamma, F, \vec{f} \rangle \in W^c$ , if  $\widehat{\mathcal{K}_i} \varphi \in \Gamma$  then there exists a  $\langle \Delta, G, \vec{g} \rangle \in W^c$  such that  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$  and  $\varphi \in \Delta$ .

PROOF Suppose  $\widehat{\mathcal{K}_i} \varphi \in \Gamma$ . we will construct a  $\langle \Delta, G, \vec{g} \rangle$  such that  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$  and  $\varphi \in \Delta$ . Let  $\Delta^-$  be  $\{\varphi\} \cup \{\mathcal{K}y_i \psi \mid \mathcal{K}y_i \psi \in \Gamma\} \cup \{\chi \mid \mathcal{K}_i \chi \in \Gamma\}$ . Then  $\Delta^-$  is consistent. Suppose not, then there are  $\mathcal{K}y_i \psi_1, \dots, \mathcal{K}y_i \psi_m, \chi_1, \dots, \chi_n \in \Delta^-$  such that  $\vdash_{\text{SKY}} \mathcal{K}y_i \psi_1 \wedge \dots \wedge \mathcal{K}y_i \psi_m \wedge \chi_1 \wedge \dots \wedge \chi_n \rightarrow \neg \varphi$ . Then  $\vdash_{\text{SKY}} \mathcal{K}_i(\mathcal{K}y_i \psi_1 \wedge \dots \wedge \mathcal{K}y_i \psi_m \wedge \chi_1 \wedge \dots \wedge \chi_n) \rightarrow \mathcal{K}_i \neg \varphi$ . Since  $\vdash_{\text{SKY}} (\mathcal{K}_i \mathcal{K}y_i \psi_1 \wedge \dots \wedge \mathcal{K}_i \mathcal{K}y_i \psi_m \wedge \mathcal{K}_i \chi_1 \wedge \dots \wedge \mathcal{K}_i \chi_n) \rightarrow \mathcal{K}_i(\mathcal{K}y_i \psi_1 \wedge \dots \wedge \mathcal{K}y_i \psi_m \wedge \chi_1 \wedge \dots \wedge \chi_n)$ , hence by the propositional calculus,  $\vdash_{\text{SKY}} (\mathcal{K}_i \mathcal{K}y_i \psi_1 \wedge \dots \wedge \mathcal{K}_i \mathcal{K}y_i \psi_m \wedge \mathcal{K}_i \chi_1 \wedge \dots \wedge \mathcal{K}_i \chi_n) \rightarrow \mathcal{K}_i \neg \varphi$ . By  $\mathcal{K}y_i \psi_j \in \Gamma$  and axiom 4YK, we have  $\mathcal{K}_i \mathcal{K}y_i \psi_j \in \Gamma$ . Since  $\mathcal{K}_i \chi_j \in \Gamma$ , it

follows that  $\mathcal{K}_i \neg \varphi \in \Gamma$ , i.e.,  $\neg \widehat{\mathcal{K}_i} \varphi \in \Gamma$ . But this is impossible:  $\Gamma$  is an MCS containing  $\widehat{\mathcal{K}_i} \varphi$ . We conclude that  $\Delta^-$  is consistent.

Let  $\Delta$  be any MCS containing  $\Delta^-$ , such extensions exist by a Lindenbaum-like argument. It follows that for any  $\mathcal{K}_{y_i} \varphi$ ,  $\mathcal{K}_{y_i} \varphi \in \Gamma$  iff  $\mathcal{K}_{y_i} \varphi \in \Delta$ :

- Suppose  $\mathcal{K}_{y_i} \varphi \in \Gamma$ . By the construction of  $\Delta$ , we have  $\mathcal{K}_{y_i} \varphi \in \Delta$ .
- Suppose  $\mathcal{K}_{y_i} \varphi \in \Delta$  and  $\mathcal{K}_{y_i} \varphi \notin \Gamma$ . By the property of MCS, we have  $\neg \mathcal{K}_{y_i} \varphi \in \Gamma$ . By axiom 5YK, we have  $\mathcal{K}_i \neg \mathcal{K}_{y_i} \varphi \in \Gamma$ . By the construction of  $\Delta$ , we have  $\neg \mathcal{K}_{y_i} \varphi \in \Delta$ . Contradiction.

In the following, we construct  $G$  and  $\vec{g}$  to form a world  $\langle \Delta, G, \vec{g} \rangle$  in  $W^c$ . Based on the above result, we can simply let  $g_i = f_i$  since  $\text{dom}(f_i) = \text{dom}(g_i)$  and let  $F \subseteq G$ . We just need to construct  $g_j$  for  $j \neq i$ . Formally, let:

- $G_0 = F \cup \{ \langle \varphi, \varphi \rangle \mid \mathcal{K}_{y_j} \varphi \in \Delta \text{ for some } j \neq i \}$
- $G_{n+1} = G_n \cup \{ \langle s \cdot t, \psi \rangle \mid \langle s, \varphi \rightarrow \psi \rangle, \langle t, \varphi \rangle \in G_n \}$
- $G = \bigcup_{n \in \mathbb{N}} G_n$

$$g_j(\varphi) = \begin{cases} f_j(\varphi) & j = i, \\ \varphi & j \neq i. \end{cases}$$

Since  $F \subseteq G$  and  $G$  is closed under implication thus conditions (i) and (ii) are obvious. For condition (iii), if  $\mathcal{K}_{y_i} \varphi \in \Delta$  then  $\mathcal{K}_{y_i} \varphi \in \Gamma$  thus  $\langle g_i(\varphi), \varphi \rangle = \langle f_i(\varphi), \varphi \rangle \in F \subseteq G$ . Condition (iii) also holds if  $\mathcal{K}_{y_j} \varphi \in \Delta$  for  $j \neq i$  by definition of  $G_0$ . It follows that  $\langle \Delta, G, \vec{g} \rangle \in W^c$ . By the construction of  $\langle \Delta, G, \vec{g} \rangle$ , we have  $\varphi \in \Delta$ ,  $\Gamma_i^\# \subseteq \Delta$ , and  $f_i = g_i$ . Therefore there exists a state  $\langle \Delta, G, \vec{g} \rangle \in W^c$  such that  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$  and  $\varphi \in \Delta$ .  $\square$

To refute  $\mathcal{K}_{y_i} \psi$  semantically, for each explanation  $t$  for  $\psi$  at the current world, we need to construct an accessible world where  $t$  is not an explanation for  $\psi$ . This leads to the following lemma.

**Lemma 20 ( $\mathcal{K}_{y_i}$  Existence Lemma)** *For any  $\langle \Gamma, F, \vec{f} \rangle \in W^c$ , if  $\mathcal{K}_{y_i} \psi \notin \Gamma$  then for any  $\langle t, \psi \rangle \in F$ , there exists  $\langle \Delta, G, \vec{g} \rangle \in W^c$  such that  $\langle t, \psi \rangle \notin G$  and  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$ .*

**PROOF** Suppose  $\mathcal{K}_{y_i} \psi \notin \Gamma$ ,  $\langle \Gamma, F, \vec{f} \rangle \in W^c$ , and  $\langle t, \psi \rangle \in F$ . We construct  $\langle \Delta, G, \vec{g} \rangle$  as follows:

- $\Delta = \Gamma$
- $\Psi = \{ \langle s, \varphi \rangle \mid \langle s, \varphi \rangle \in F \text{ and } \mathcal{K}_{y_i} \varphi \notin \Gamma \}$
- $\Psi' = \{ \langle t \cdot s, \varphi \rangle \mid \langle s, \varphi \rangle \in \Psi \}$

- $G_0 = (F \setminus \Psi) \cup \Psi'$
- $G_{n+1} = G_n \cup \{\langle r \cdot s, \varphi_2 \rangle \mid \langle r, \varphi_1 \rightarrow \varphi_2 \rangle, \langle s, \varphi_1 \rangle \in G_n\}$
- $G = \bigcup_{n \in \mathbb{N}} G_n$
- For each  $j \in \mathbf{I}$ ,  $g_j : \{\varphi \mid \mathcal{K}y_j\varphi \in \Delta\} \rightarrow E^c$  is defined as:

$$g_j(\varphi) = \begin{cases} f_j(\varphi), & \langle f_j(\varphi), \varphi \rangle \notin \Psi \\ t \cdot f_j(\varphi), & \langle f_j(\varphi), \varphi \rangle \in \Psi \end{cases} \quad (1)$$

From the construction of  $G$ , it is clear that for any  $\langle s, \varphi \rangle \in \Psi'$ ,  $|s| > |t|$ , thus in particular  $\langle t, \psi \rangle$  is not in  $G_0$ . The idea behind the construction of  $G$  is to first replace any current explanation for  $\psi$  with something longer, and then take the closure w.r.t. implication. Note that for technical convenience, we treat all  $\varphi$  such that  $\mathcal{K}y_i\varphi \notin \Gamma$  in the basic step together.

Now we prove the following claims.

**Claim 1**  $\langle \Delta, G, \vec{g} \rangle \in W^c$ . i.e.,  $G$  satisfies the conditions in the definition of  $W^c$ .

- (i) Suppose  $\langle r, \varphi_1 \rightarrow \varphi_2 \rangle, \langle s, \varphi_1 \rangle \in G$ . By the construction of  $G$ , there exists  $n \in \mathbb{N}$  such that  $\langle r, \varphi_1 \rightarrow \varphi_2 \rangle, \langle s, \varphi_1 \rangle \in G_n$ . By the construction of  $G_{n+1}$ , we have  $\langle r \cdot s, \varphi_2 \rangle \in G_{n+1}$ . Thus  $\langle r \cdot s, \varphi_2 \rangle \in G$ .
- (ii) Suppose  $\varphi \in \Lambda$ . Then  $\langle e, \varphi \rangle \in F$ . Since  $\varphi$  is a tautology, by NECKY and the property of MCS, we have  $\mathcal{K}y_i\varphi \in \Gamma$ . Thus  $\langle e, \varphi \rangle \notin \Psi$ . Thus  $\langle e, \varphi \rangle \in G_0$ . Hence  $\langle e, \varphi \rangle \in G$ .
- (iii) Suppose  $\mathcal{K}y_j\varphi \in \Delta$  ( $j \in \mathbf{I}$ ). Since  $\Delta = \Gamma$ , we have  $\mathcal{K}y_j\varphi \in \Gamma$ . Thus  $\langle f_j(\varphi), \varphi \rangle \in F$ . We have two cases:
  - $\langle f_j(\varphi), \varphi \rangle \notin \Psi$ : Thus  $g_j(\varphi) = f_j(\varphi)$ . Thus  $\langle g_j(\varphi), \varphi \rangle \in F$  and  $\langle g_j(\varphi), \varphi \rangle \notin \Psi$ . Thus  $\langle g_j(\varphi), \varphi \rangle \in G_0$ . Hence  $\langle g_j(\varphi), \varphi \rangle \in G$ .
  - $\langle f_j(\varphi), \varphi \rangle \in \Psi$ : Thus  $g_j(\varphi) = t \cdot f_j(\varphi)$  and  $\langle g_j(\varphi), \varphi \rangle \in \Psi'$ . Thus  $\langle g_j(\varphi), \varphi \rangle \in G_0$ . Hence  $\langle g_j(\varphi), \varphi \rangle \in G$ .

**Claim 2**  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$

To prove this claim, we just need to check two conditions:

- (1) Since  $\Delta = \Gamma$ , obviously, we have  $\Gamma_i^\# \subseteq \Delta$ .
- (2) Since  $\Delta = \Gamma$ , thus  $\{\varphi \mid \mathcal{K}y_i\varphi \in \Gamma\} = \{\varphi \mid \mathcal{K}y_i\varphi \in \Delta\}$ . i.e.,  $\text{dom}(g_i) = \text{dom}(f_i)$ . For any  $\varphi \in \{\varphi \mid \mathcal{K}y_i\varphi \in \Delta\}$ , since  $\langle f_i(\varphi), \varphi \rangle \notin \Psi$ , by the definition of  $g_i$ , we have  $g_i(\varphi) = f_i(\varphi)$ . Hence  $g_i = f_i$ .

To prove  $\langle t, \psi \rangle \notin G$ , we first prove the following useful claim:

**Claim 3** If  $\mathcal{K}y_i\varphi \notin \Gamma$  and  $\langle s, \varphi \rangle \in G_{n+1} \setminus G_n$ , then  $|s| > |t|$ .

Suppose  $\mathcal{K}y_i\varphi \notin \Gamma$ . Do induction on  $n$ :



- $n = 0$ . Suppose  $\langle s, \varphi \rangle \in G_1 \setminus G_0$ . Then there exists  $s_1, s_2$ , and  $\chi$  such that  $s = s_1 \cdot s_2$ ,  $\langle s_1, \chi \rightarrow \varphi \rangle, \langle s_2, \chi \rangle \in G_0$ . We have two cases
  - $\langle s_1, \chi \rightarrow \varphi \rangle \in \Psi'$  or  $\langle s_2, \chi \rangle \in \Psi'$ : Thus  $|s_1| > |t|$  or  $|s_2| > |t|$ . Thus  $|s| > |t|$ .
  - $\langle s_1, \chi \rightarrow \varphi \rangle \notin \Psi'$  and  $\langle s_2, \chi \rangle \notin \Psi'$ : Since  $\langle s_1, \chi \rightarrow \varphi \rangle, \langle s_2, \chi \rangle \in G_0$ , thus  $\langle s_1, \chi \rightarrow \varphi \rangle, \langle s_2, \chi \rangle \in F \setminus \Psi$ . Thus  $\mathcal{K}y_i(\chi \rightarrow \varphi), \mathcal{K}y_i\chi \in \Gamma$ . Thus  $\mathcal{K}y_i\varphi \in \Gamma$  by axiom DISTY. Contradiction.
- Induction Hypothesis: If  $\langle s, \varphi \rangle \in G_{k+1} \setminus G_k$ , then  $|s| > |t|$ . Suppose  $\langle s, \varphi \rangle \in G_{k+2} \setminus G_{k+1}$ . Then we have  $\langle s, \varphi \rangle \in \{\langle s_1 \cdot s_2, \varphi_2 \rangle \mid \langle s_1, \varphi_1 \rightarrow \varphi_2 \rangle, \langle s_2, \varphi_1 \rangle \in G_{k+1}\}$ . Thus there exists  $s_1, s_2, \chi$  such that  $s = s_1 \cdot s_2$ ,  $\langle s_1, \chi \rightarrow \varphi \rangle, \langle s_2, \chi \rangle \in G_{k+1}$ . Since  $\langle s, \varphi \rangle \notin G_{k+1}$ , then we have  $\langle s_1, \chi \rightarrow \varphi \rangle \notin G_k$  or  $\langle s_2, \chi \rangle \notin G_k$ . Thus  $\langle s_1, \chi \rightarrow \varphi \rangle \in G_{k+1} \setminus G_k$  or  $\langle s_2, \chi \rangle \in G_{k+1} \setminus G_k$ . By IH, we have  $|s_1| > |t|$  or  $|s_2| > |t|$ . Hence  $|s| > |t|$ .

**Claim 4**  $\langle t, \psi \rangle \notin G$ .

According to the construction of  $G$ , we just need to show that for all  $n \in \mathbb{N}$ ,  $\langle t, \psi \rangle \notin G_n$ . Based on Claim 3,  $\langle t, \psi \rangle$  cannot be added in any  $G_n$  for  $n \geq 1$ . Therefore we just need to show that  $\langle t, \psi \rangle \notin G_0$ , but this is clear due to the way we constructed  $G_0$ .  $\square$

Finally we are ready to prove the truth lemma.

**Lemma 21 (Truth Lemma)** *For all  $\varphi$ ,  $\langle \Gamma, F, \vec{f} \rangle \models \varphi$  if and only if  $\varphi \in \Gamma$ .*

**PROOF** This is established by standard induction on the complexity of  $\varphi$ . The atomic cases and the boolean cases are standard. The case when  $\varphi = \mathcal{K}_i\psi$  is also routine based on Lemma 19.

Consider the case that  $\varphi$  is  $\mathcal{K}_i\psi$  for some  $\psi$ .

- $\Leftarrow$  If  $\mathcal{K}_i\psi \in \Gamma$ , for any  $\langle \Delta, G, \vec{g} \rangle$  such that  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$ , we have then  $\mathcal{K}_i\psi \in \Delta$  by the definition of  $R_i^c$ . Since  $\vdash_{\text{SKY}} \mathcal{K}_i\psi \rightarrow \psi$  (by T and PRES), we have  $\psi \in \Delta$ . By the IH, we have  $\langle \Delta, G, \vec{g} \rangle \models \psi$ . Since  $\mathcal{K}_i\psi \in \Gamma$  and  $\mathcal{K}_i\psi \in \Delta$ , then we have  $\langle f_i(\psi), \psi \rangle \in F$  and  $\langle g_i(\psi), \psi \rangle \in G$ . By the definition of  $R_i^c$ , we have  $f_i = g_i$ . Thus there exists  $g_i(\psi) = f_i(\psi) \in E^c$  such that  $\langle \Delta, G, \vec{g} \rangle \in \mathcal{E}^c(g_i(\psi), \psi)$ . Therefore  $\langle \Gamma, F, \vec{f} \rangle \models \mathcal{K}_i\psi$ .
- $\Rightarrow$  Suppose  $\mathcal{K}_i\psi \notin \Gamma$ . We have two cases as follows:
  - $\mathcal{K}_i\psi \notin \Gamma$ : then by Lemma 19 and the semantics,  $\langle \Gamma, F, \vec{f} \rangle \not\models \mathcal{K}_i\psi$ .
  - $\mathcal{K}\psi \in \Gamma$ : We also have two cases:
    - \*  $\langle t, \psi \rangle \notin F$  for all  $t \in E$ . By the semantics,  $\langle \Gamma, F, \vec{f} \rangle \not\models \mathcal{K}_i\psi$ .

- \* There exists  $t \in E$  such that  $\langle t, \psi \rangle \in F$ . By Lemma 20, we have that for any  $\langle t, \psi \rangle \in F$ , there exists  $\langle \Delta, G, \vec{g} \rangle \in W^c$  such that  $\langle t, \psi \rangle \notin G$  and  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$ . Hence we have  $\langle \Gamma, F, \vec{f} \rangle \not\models \mathcal{K}y_i \psi$ .

□

**Theorem 22 (Completeness of SKY over  $\mathbb{C}$ )**  $\Sigma \models_{\mathbb{C}} \varphi$  implies  $\Sigma \vdash_{\text{SKY}} \varphi$ .

PROOF Suppose  $\Sigma \models_{\mathbb{C}} \varphi$ . Towards contradiction, suppose  $\Sigma \not\vdash_{\text{SKY}} \varphi$ . Then  $\Sigma \cup \{\neg\varphi\}$  is consistent. Extend  $\Sigma \cup \{\neg\varphi\}$  to a maximal consistent set  $\Gamma$ . By Proposition 18, there exist  $F$  and  $\vec{f}$  such that  $\langle \Gamma, F, \vec{f} \rangle \in W^c$ . By Lemma 21, we have  $\langle \Gamma, F, \vec{f} \rangle \models \Sigma \cup \{\neg\varphi\}$ , thus  $\Sigma \cup \{\neg\varphi\}$  is satisfiable, thus  $\Sigma \models_{\mathbb{C}} \varphi$  is false. Contradiction. □

By theorem 8 and theorem 22, we have the following corollary.

**Corollary 23 (Completeness of SKY over  $\mathbb{C}_F$ )**  $\Sigma \models_{\mathbb{C}_F} \varphi$  implies  $\Sigma \vdash_{\text{SKY}} \varphi$ .

Now let us look at the completeness of SKYI. The crucial observation is that we can use the same canonical model definition except now we let  $\Omega$  be the set of all maximal SKYI-consistent set of **ELK<sub>y</sub>** formulas. The similar propositions follow due to Proposition 11. The only extra thing is to check whether the new canonical model has the introspection property.

**Proposition 24**  $\mathcal{M}^c$  has introspection property.

PROOF Suppose  $\langle \Gamma, F, \vec{f} \rangle \models \varphi$  and  $\varphi$  has the form of  $\mathcal{K}_i \psi$  or  $\neg \mathcal{K}_i \psi$  or  $\mathcal{K}y_i \psi$  or  $\neg \mathcal{K}y_i \psi$ . By Lemma 21, we have  $\varphi \in \Gamma$ . By the axioms 4KY-5Y and the properties of MCS, we have  $\mathcal{K}y_i \varphi \in \Gamma$ . By Lemma 21, we have  $\langle \Gamma, F, \vec{f} \rangle \models \mathcal{K}y_i \varphi$ . Thus  $\exists r \in E^c$ ,  $\langle \Delta, G, \vec{g} \rangle \in \mathcal{E}^c(r, \varphi)$  for each  $\langle \Delta, G, \vec{g} \rangle$  such that  $\langle \Gamma, F, \vec{f} \rangle R_i^c \langle \Delta, G, \vec{g} \rangle$ . □

Based on the above proposition and Theorem 10 we have:

**Theorem 25 (Completeness of SKYI over  $\mathbb{C}_I$  and  $\mathbb{C}_{FI}$ )** If  $\Sigma \models_{\mathbb{C}_I} \varphi$ , then  $\Sigma \vdash_{\text{SKYI}} \varphi$ . If  $\Sigma \models_{\mathbb{C}_{FI}} \varphi$ , then  $\Sigma \vdash_{\text{SKYI}} \varphi$ .

## 4 Comparison with justification logic

In this section, we compare our framework with justification logic.

The language of the most classic justification logic LP (i.e., JT4 in [3]) includes both formulas  $\varphi$  and justification terms  $t$ :

$$\begin{aligned}\varphi &::= \top \mid p \mid \neg\varphi \mid (\varphi \wedge \varphi) \mid t:\varphi \\ t &::= x \mid c \mid (t \cdot t) \mid (t + t) \mid !t\end{aligned}$$

The possible-world semantics of justification logic is based on the Fitting model  $\langle S, R, \mathcal{E}, V \rangle$  where  $\langle S, R, V \rangle$  is a single-agent Kripke model and  $\mathcal{E}$  is an evidence function assigning justification terms  $t$  to formulas on each world, just as in our setting. The formula  $t:\varphi$  has the following semantics (cf. e.g., [10]):

$\begin{aligned}\mathcal{M}, w \Vdash t:\varphi &\iff \text{(a) } w \in \mathcal{E}(t, \varphi); \\ &\text{(b) } v \Vdash \varphi \text{ for all } v \text{ such that } wRv.\end{aligned}$
--

Compared to our semantics for  $\mathcal{K}y_i\phi$ , note that (a) only requires that  $t$  is a justification of  $\phi$  on the current world  $w$ . The Fitting models for LP are assumed to have further conditions:<sup>10</sup>

1.  $\mathcal{E}(s, \varphi \rightarrow \psi) \cap \mathcal{E}(t, \varphi) \subseteq \mathcal{E}(s \cdot t, \psi)$
2.  $\mathcal{E}(t, \varphi) \cup \mathcal{E}(s, \varphi) \subseteq \mathcal{E}(s + t, \varphi)$
3.  $\mathcal{E}(t, \varphi) \subseteq \mathcal{E}(!t, t:\varphi)$
4. Monotonicity:  $w \in \mathcal{E}(t, \varphi)$  and  $wRv$  implies  $v \in \mathcal{E}(t, \varphi)$ .
5.  $R$  is reflexive and transitive.

Note that we also require (1) and (5) above and include  $\cdot$  as an operation on explanations in  $E$  semantically. On the other hand, we leave out (2)(3)(4) and the operations  $+$  and  $!$  for specific considerations in our setting. For the case of  $+$ , consider the following model where  $\varphi$  has two possible explanations and agent  $i$  cannot distinguish them (thus  $\neg\mathcal{K}y_i\varphi$  holds).

$$t:\varphi \text{ --- } i \text{ --- } s:\varphi$$

If we impose condition (2) then  $s + t$  is a uniform explanation of  $\varphi$  on both worlds, which makes  $\mathcal{K}y_i\varphi$  true. More generally, for any finite model where  $\varphi$  has some explanations on each world,  $\mathcal{K}y_i\varphi$  will always be true under condition (2), which is counterintuitive in our setting. Conceptually, the explanation should be *precise*, you cannot explain a theorem by saying one of all the possible proofs up to a certain length should work. Knowing there is a proof does not mean you know why the theorem holds.

<sup>10</sup> The “S5 version” of justification logic JT45 also adds another condition about negative introspection:  $\mathcal{E}(t, \varphi) \subseteq \mathcal{E}(!t, \neg(t:\varphi))$ , and requires strong evidence, where  $!$  is a new operation for justification terms in the language, cf. [3].

Operation  $!$  and conditions (3) and (4) are relevant to the validity of the axiom  $t:\varphi \rightarrow !t:(t:\varphi)$  in justification logic LP, which is used to realize axiom 4 in modal logic. Intuitively,  $!$  is the proof checker and  $!t$  can always be a justification of  $t:\phi$ .<sup>11</sup> Although we do not have  $t:\varphi$  in the language, it may sound reasonable to include  $!$  and require  $\mathcal{E}(t, \varphi) \subseteq \mathcal{E}(!t, \mathcal{K}_{y_i}\varphi)$  instead. However, due to the desired factivity ( $w \in \mathcal{E}(t, \varphi)$  implies  $w \models \varphi$ ),  $\mathcal{K}_{y_i}\varphi$  will then hold on each world where  $\varphi$  has an explanation, which is counterintuitive. Conceptually, that  $t$  is an explanation for  $\varphi$  does not entail  $t$  can be transformed uniformly into an explanation for  $\mathcal{K}_{y_i}\varphi$ . For example, the window is broken since someone threw a rock at it, but there can be different explanations for an agent to know why the window is broken: she saw it, or someone told her about it, and so on. The technically motivated condition (4) in justification logic intuitively requires that one can only imagine more explanations than the real world which is also not reasonable in our setting, as an undesired consequence will follow:  $w \in \mathcal{E}(t, \varphi)$  and  $w \models \mathcal{K}_i\varphi$  imply  $w \models \mathcal{K}_{y_i}\varphi$  due to condition (4).

In [4,2], languages with both  $\Box_i$  and  $t:\varphi$  formulas are discussed. In the semantics, an extra evidence accessibility relation  $R^e$  is introduced to interpret  $t:\varphi$  while  $\Box_i\varphi$  is interpreted by  $R_i$ . In [2],  $t:\varphi$  says that  $\varphi$  is justified common knowledge, and it is required that  $(\bigcup_{i \in I} R_i)^+ \subseteq R^e$  in order to make the axiom  $t:\varphi \rightarrow \Box_i\varphi$  valid. In [4], it is required that  $R_i \subseteq R^e$ .<sup>12</sup> Conceptually, the requirement is based on the idea that you may implicitly know more than those that are justified. In our work, we do not assume there is an extra (objective) evidence accessibility relation  $R^e$ . On the other hand, we do have the analogous axiom  $\mathcal{K}_{y_i}\varphi \rightarrow \mathcal{K}_i\varphi$ . In some works on multi-agent justification logic, the evidence function  $\mathcal{E}$  is agent-dependent: for each agent and each world some justification is given to some formulas. Here we assume the explanation function is independent from the agents.

In justification logic, there is always a constant specification (CS), a collection of  $c_1 : c_2 : \dots c_n : \varphi$  formulas where  $\varphi$  is an axiom in the corresponding logic. It makes sure that all the axioms in the corresponding logic are justified by special constants in any depth, thus can be used in the reasoning in the proof system. A model meets the requirement of a CS if  $W = \mathcal{E}(t, \varphi)$  for all  $t:\varphi \in CS$ . In contrast, we do not include all the axioms in our tautology ground  $\Lambda$  on purpose. For example, if  $T: (\mathcal{K}_i\varphi \rightarrow \varphi) \in \Lambda$  then we have  $\mathcal{K}_{y_i}(\mathcal{K}_i\varphi \rightarrow \varphi)$  by NECKY. It follows that

<sup>11</sup> In the multi-agent setting,  $!$  was introduced to capture the proof check done by each agent [35].

<sup>12</sup> Though in a single agent setting.

$\mathcal{K}y_i\mathcal{K}_i\varphi \rightarrow \mathcal{K}y_i\varphi$  by DISTY, which may sound strange: e.g., I know why I know that the window is broken implies I know why it is broken.

The table below highlights the similarities between our axioms (or derivable theorems in SKY and SKYI) and axioms in (variants of) justification logic when viewing  $t:\phi$  as  $\mathcal{K}y_i\phi$ :

Justification Logic	Our work
$t:(\varphi \rightarrow \psi) \rightarrow s:\varphi \rightarrow (t \cdot s):\psi$	$\mathcal{K}y_i(\varphi \rightarrow \psi) \rightarrow (\mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\psi)$
$t:\varphi \rightarrow (s + t):\varphi$	$\mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\varphi$
$t:\varphi \rightarrow !t:(t:\varphi)$	$\mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\mathcal{K}y_i\varphi$
$\neg t:\varphi \rightarrow ?t:(\neg t:\varphi)$	$\neg \mathcal{K}y_i\varphi \rightarrow \mathcal{K}y_i\neg \mathcal{K}y_i\varphi$
$t:\varphi \rightarrow \varphi$	$\mathcal{K}y_i\varphi \rightarrow \varphi$
$t:\varphi \rightarrow \Box\varphi$ [4]	$\mathcal{K}y_i\varphi \rightarrow \mathcal{K}_i\varphi$
$t:\varphi \rightarrow \Box t:\varphi$ [4]	$\mathcal{K}y_i\varphi \rightarrow \mathcal{K}_i\mathcal{K}y_i\varphi$
$\neg t:\varphi \rightarrow \Box\neg t:\varphi$ [4]	$\neg \mathcal{K}y_i\varphi \rightarrow \mathcal{K}_i\neg \mathcal{K}y_i\varphi$

Based on the above analogy, it seems that our logic is partially realizable in S5(JT45), the logic with S5- $\Box$  and JT45- $t:\phi$ .<sup>13</sup> As we mentioned, our language can also be viewed as a fragment of the quantified justification logic proposed in [9] extended with  $\Box_i$  modalities. However, the domain for justifications is not assumed to be constant in the FO models of [9].

## 5 Conclusions and Future work

In this paper, we present an attempt to formalize the logic of knowing why. In the language we have both the standard knowing that operator  $\mathcal{K}_i$  and the new knowing why operator  $\mathcal{K}y_i$ . A semantics based on Fitting models for justification logic is given, which interprets knowing why  $\phi$  as there exists an explanation such that I know it is one explanation for  $\phi$ . We gave two proof systems, one weaker and one stronger depending on the choice of introspection axioms, and showed their completeness over various model classes.

As the title shows, it is by no means *the* logic of knowing why. Besides the introspection axioms, there are a lot to be discussed.<sup>14</sup> For example, although DISTY looks reasonable in a setting focusing on deductive explanations, it may cause troubles if causal explanations or other types of explanations are considered. Recall our example about the flagpole and its shadow. It is reasonable to assume that I know why the shadow is  $y$

<sup>13</sup> Note that we do not have the CS as in justification logic but a tautology ground  $\Lambda$ . This may cause trouble to the realization result.

<sup>14</sup> We may also think whether  $\mathcal{K}y_i\phi \rightarrow \mathcal{K}y_i\mathcal{K}_i\phi$  is reasonable.

meters long ( $\mathcal{K}_{y_i}p$ ), and I also know why that the shadow is  $y$  meters implies the pole is  $x$  meters long ( $\mathcal{K}_{y_i}(p \rightarrow q)$ ). However, it does not entail that I know why the pole is  $x$  meters long ( $\mathcal{K}_{y_i}q$ ) if we are looking for causal explanation (or functional explanation). One way to go around is to replace the material implication by some relevant (causal) conditionals, then  $\mathcal{K}_{y_i}(p \rightarrow q)$  may not be there anymore.

It seems we often do not have clear semantic intuition about non-trivial expressions of knowing why. One reason is that there may be different readings of the same statement of knowing why  $\phi$  regarding different aspects of  $\phi$  and different types of desired explanations. For example, “I know why Frank went to Beijing on Monday” may have different meanings depending on the *contrast* the speaker wants to emphasize [30]:

- I know why *Frank*, not Mary, went to Beijing on Monday.
- I know why Frank went to *Beijing*, not Shanghai, on Monday.
- I know why Frank went to Beijing on *Monday*, not on Tuesday.

Following [20], we may partially handle this by adding contrast formulas, e.g., turn  $\mathcal{K}_{y_i}\varphi$  into  $\mathcal{K}_{y_i}(\varphi \wedge \neg\psi \wedge \neg\chi)$  depending on the emphasis. However, we cannot handle the changes of types of explanations depending on the contrast.

It is also interesting to study alternative semantics of  $\mathcal{K}_i$  in our setting. Thomas Studer proposed the following semantics for  $\mathcal{K}_i$ :<sup>15</sup>

$$\boxed{\mathcal{M}, w \models \mathcal{K}_i\varphi \iff \text{for each } v \text{ s.t. } wR_iv : v \models \varphi \text{ and } v \in \mathcal{E}(t, \varphi) \text{ for some } t.}$$

Then the main distinction between  $\mathcal{K}_i\varphi$  and  $\mathcal{K}_{y_i}\varphi$  becomes the distinction of *de dicto* and *de re* readings of knowing why:  $\mathcal{K}_i\exists t(t: \varphi)$  and  $\exists t\mathcal{K}_i(t: \varphi)$ . This semantics is clearly more demanding for  $\mathcal{K}_i$  as you need to have some explanation, and it will affect the logic. For example,  $\mathcal{K}_i\varphi \rightarrow \mathcal{K}_i\mathcal{K}_i\varphi$  no longer holds on S5 models. However, if we consider only introspective models and include T in  $\Delta$ , then the two semantics for  $\mathcal{K}_i$  coincide.<sup>16</sup>

Another future direction is to study the inner structure of explanations further. Hintikka’s early work [18] may turn out to be helpful, where explanations can be of the form of a universally quantified formula, which connects better with the existing theories of scientific explanations in philosophy of science.<sup>17</sup> Moreover, we may be interested in saying an explanation is true. The factivity that we proposed did not fully capture that.

<sup>15</sup> via personal communication.

<sup>16</sup> By such conditions, we can rule out the case that you know  $\varphi$  without an explanation.

<sup>17</sup> There are also modal logic approaches to handle scientific explanations cf. e.g., [28, 27].

A promising future study is about group notions of knowing why. For example, how do we define everyone knows why  $\varphi$ ? Simply having a conjunction of  $\mathcal{K}_{y_i}\varphi$  for each  $i$  may not be enough, since people can have different explanations for  $\varphi$ . The case of *commonly* knowing why  $\varphi$  is more interesting. For example, we may have different definitions:

- It is (standard) common knowledge that everyone knows why  $\varphi$  w.r.t. the same explanation.
- Everyone knows why ... everyone knows why  $\varphi$ .

In contrast to standard epistemic logic, such definitions can be quite different from each other. Since each iteration of  $\mathcal{K}_{y_i}$  may ask for a new explanation, we then have a much richer spectrum of such common knowledge notions, e.g., for the second definition, we may ask the agents to have exactly the same explanation for each level of “iteration of everyone knows why”. It will be interesting to compare such notions with justified common knowledge proposed in [2].

Of course, we can also consider the dynamics of knowing why, similar to the dynamics in justification logic [21,22]. Clearly, public announcements can change knowledge-why but there can be more natural dynamics, e.g., publicly announcing why, which is similar to public inspection introduced in [11] in the setting of knowing values. A deeper connection between knowing why and dynamic epistemic logic can be established based on the observation that we do update according to events because we know why they happened (the preconditions). It is suggested by Olivier Roy that there is also a close connection with forward induction in games, where it is crucial to guess why someone did an apparently irrational move.

Finally, our work is also related to *explicit knowledge*, which aims to avoid logical omniscience. In fact, knowledge with justification or explanation can be viewed as a type of explicit knowledge. One important approach to define explicit knowledge is by using *awareness*:  $\varphi$  is a piece of explicit knowledge of  $i$  ( $X_i\varphi$ ) if  $i$  is aware of  $\varphi$  ( $A_i\varphi$ ) and  $i$  implicitly knows that  $\varphi$  ( $\mathcal{K}_i\varphi$ ), where awareness is often defined syntactically (cf. [7]). Accordingly, the axioms are also changed, e.g., the K axiom now becomes  $X_i(\varphi \rightarrow \psi) \wedge X_i\varphi \wedge A_i\psi \rightarrow X_i\psi$ . Other approaches to explicit knowledge uses idea of *algorithmic knowledge* [12]. We may explore the concrete connection in the future.

*Acknowledgement* We thank Albert Anglberger, Huimin Dong, Mel Fitting, Domink Klein, Fenrong Liu, Olivier Roy, Thomas Studer, Wei Wang, Junhua Yu, Jun Zhang, and Liying Zhang for useful suggestions on earlier



versions of this paper. Yanjing Wang acknowledges the support from the National Program for Special Support of Eminent Professionals and NSSF key projects 12&ZD119.

## References

1. Sergei Artemov. Operational modal logic, 1995.
2. Sergei Artemov. Justified common knowledge. *Theoretical Computer Science*, 357(1):4 – 22, 2006.
3. Sergei Artemov. The logic of justification. *The Review of Symbolic Logic*, 1(04):477–513, 2008.
4. Sergei N. Artëmov and Elena Nogina. Introducing justification into epistemic logic. *Journal of Logic and Computation*, 15(6):1059–1073, 2005.
5. Alexander Bird. *Philosophy of science*. Routledge, 1998.
6. Sylvain Bromberger. Questions. *The Journal of Philosophy*, 63(20):597–606, 1966.
7. R. Fagin, J. Halpern, Y. Moses, and M. Vardi. *Reasoning about knowledge*. MIT Press, 1995.
8. Melvin Fitting. The logic of proofs, semantically. *Annals of Pure and Applied Logic*, 132(1):1–25, 2005.
9. Melvin Fitting. A quantified logic of evidence. *Annals of Pure and Applied Logic*, 152(1):67 – 83, 2008.
10. Melvin Fitting. Modal logics, justification logics, and realization. *Ann. Pure Appl. Logic*, 167(8):615–648, 2016.
11. Malvin Gattinger, Jan van Eijck, and Yanjing Wang. Knowing values and public inspection. under submission, 2016.
12. J. Y. Halpern and R. Pucella. Dealing with logical omniscience: Expressiveness and pragmatics. *Artificial Intelligence*, 175(1):220–235, 2011.
13. Carl Hempel. *Aspects of Scientific Explanation and Other Essays in the Philosophy of Science*. The Free Press, 1965.
14. Carl G Hempel and Paul Oppenheim. Studies in the logic of explanation. *Philosophy of science*, 15(2):135–175, 1948.
15. Jaakko Hintikka. *Knowledge and belief: an introduction to the logic of the two notions*, volume 181. Cornell University Press Ithaca, 1962.
16. Jaakko Hintikka. On the logic of an interrogative model of scientific inquiry. *Synthese*, 47(1):69–83, 1981.
17. Jaakko Hintikka. *New Foundations for a Theory of Questions and Answers*, pages 159–190. Springer Netherlands, Dordrecht, 1983.
18. Jaakko Hintikka and Ilpo Halonen. Semantics and pragmatics for why-questions. *The Journal of Philosophy*, 92(12):636–657, 1995.
19. Philip Kitcher. Explanatory unification. *Philosophy of Science*, 48(4):507–531, 1981.
20. Antti Koura. An approach to why-questions. *Synthese*, 74(2):191–206, 1988.
21. Bryan Renne. *Dynamic Epistemic Logic with Justification*. PhD thesis, New York, NY, USA, 2008. AAI3310607.
22. Bryan Renne. Multi-agent justification logic: communication and evidence elimination. *Synthese*, 185(1):43–82, 2012.
23. Wesley Salmon. *Scientific Explanation and the Causal Structure of the World*. Princeton University Press, 1984.
24. Gerhard Schurz. Scientific explanation: A critical survey. *Foundations of Science*, 1(3):429–465, 1995.



25. Gerhard Schurz. Explanation as unification. *Synthese*, 120(1):95–114, 1999.
26. Gerhard Schurz. Explanations in science and the logic of why-questions: Discussion of the Halonen–Hintikka-approach and alternative proposal. *Synthese*, 143(1):149–178, 2005.
27. Igor Sedlár and Juraj Halas. Modal logics of abstract explanation frameworks. Abstract in Proceedings of CLMPS 15, 2015.
28. Dunja Šešelja and Christian Straßer. Abstract argumentation and explanation applied to scientific debates. *Synthese*, 190(12):2195–2217, 2013.
29. Hans van Ditmarsch, Joseph Y. Halpern, Wiebe van der Hoek, and Barteld Kooi, editors. *Handbook of Epistemic Logic*. College Publications, 2015.
30. Bas C van Fraassen. *The scientific image*. Oxford University Press, 1980.
31. Yanjing Wang. A logic of knowing how. In *Proceedings of LORI-V*, pages 392–405, 2015.
32. Yanjing Wang. Beyond knowing that: a new generation of epistemic logics. In Hans van Ditmarsch and Gabriel Sandu, editors, *Jaakko Hintikka on knowledge and game theoretical semantics*. Springer, 2016. forthcoming.
33. Yanjing Wang and Jie Fan. Conditionally knowing what. In *Proceedings of AiML Vol. 10*, pages 569–587, 2014.
34. Erik Weber, Jeroen van Bouwel, and Leen De Vreese. *Scientific explanation*. Springer, 2013.
35. Tatiana Yavorskaya(Sidon). *Multi-agent Explicit Knowledge*, pages 369–380. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.